

Deep residual bidirectional long short-term memory fusion: achieving superior accuracy in facial emotion recognition

Muhammad Munsarif^{1,2}, Ku Ruhana Ku-Mahamud^{3,4}

¹School of Graduate Studies, Asia e University, Selangor, Malaysia

²Department of Informatics, Universitas Muhammadiyah Semarang, Semarang, Indonesia

³School of Computing, Universiti Utara Malaysia, Kedah, Malaysia

⁴Department of Information Technology, Faculty of Business, Management, and Information Technology, Universitas Muhammadiyah Malaysia, Perlis, Malaysia

Article Info

Article history:

Received Aug 1, 2024

Revised Nov 25, 2024

Accepted Dec 25, 2024

Keywords:

Bidirectional long short-term memory
Convolutional neural networks
Deep residual bidirectional long short-term memory fusion
Emotion classification
Facial expression recognition

ABSTRACT

Facial emotion recognition (FER) is a crucial task in human communication. Various face emotion recognition models were introduced but often struggle with generalization across different datasets and handling subtle variations in expressions. This study aims to develop the deep residual bidirectional long short-term memory (Bi-LSTM) fusion method to improve FER accuracy. This method combines the strengths of convolutional neural networks (CNN) for spatial feature extraction and Bi-LSTM for capturing temporal dynamics, using residual layers to address the vanishing gradient problem. Testing was performed on three face emotion datasets, and a comparison was made with seventeen models. The results show perfect accuracy on the extended Cohn-Kanade (CK+) and the real-world affective faces database (RAF-DB) datasets and almost perfect accuracy on the face expression recognition plus (FERPlus) dataset. However, the receiver operating characteristic (ROC) curve for the CK+ dataset shows some inconsistencies, indicating potential overfitting. In contrast, the ROC curves for the RAF-DB and FERPlus datasets are consistent with the high accuracy achieved. The proposed method has proven highly efficient and reliable in classifying various facial expressions, making it a robust solution for FER applications.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Muhammad Munsarif
Departement of Informatics, Universitas Muhammadiyah Semarang
St. Kedung Mundu Raya No. 18, Semarang, Central Java 50273, Indonesia
Email: m.munsarif@unimus.ac.id

1. INTRODUCTION

Facial expressions are an important part of human nonverbal communication, where facial muscle movements express emotions such as happiness, sadness, anger, surprise, fear, and disgust [1]. The ability to recognize facial expressions automatically, known as facial emotion recognition (FER), has garnered significant attention due to its potential in various applications, such as education and healthcare [2]. For instance, schools can use FER to assess teaching effectiveness, while hospitals can apply it to analyze patients' psychological conditions. Moreover, advancements in graphics processing unit (GPU) technology and the increasing use of FER across various fields have further boosted the popularity of this technology [2]. However, despite its great potential, FER faces several complex problems compared to other image classification tasks. The main problems are inter-class similarities and intra-class differences. Inter-class similarities refer to the difficulty in distinguishing between similar facial expressions, such as anger and disgust [3]. On the other hand, intra-class differences occur due to variations in expressions among individuals with different facial structures,

genders, ages, and ethnicities [4]. These variations limit how well the model generalizes, reducing its accuracy and reliability [5].

Researchers have developed various models to tackle these challenges, such as EfficientFace [6], dynamic attention network-generative adversarial network (DAN-GAN) [7], emotion-network (EMO-Net) [8], transfer learning-based facial emotion recognition (TransFER) [9], and lightweight facial emotion recognition (LiteFER) [10]. Although these models have improved FER accuracy, they still need help recognizing negative expressions effectively due to dataset imbalances and other limitations. Additionally, models like Occlusion-Aware convolutional neural networks (CNN) [11], Pseudo-3D CNN [12], and vision transformer (ViT) [13] are designed to handle challenging situations, such as when parts of the face are obscured or when the face is viewed from difficult angles. While these models provide partial solutions, further refinement is necessary to address more complex real-world scenarios. To solve these issues, researchers have suggested that using larger neural networks can improve FER performance. Models such as emotion transformer [14], DAN [15], and HybridNet [16] have shown excellent results but require significant computational power. In contrast, models like MobileEmotionNet [17], TinyFaceNet [18], Light-FER [19], multi-scale feature fusion network [20], residual masking network [21], and Context-Aware FER [22]. Pyramid attention [23], which are more lightweight and efficient, are suitable for devices with limited resources without sacrificing too much accuracy. These models provide more efficient solutions for situations with limited computational resources [17]-[23]. In addition, researchers have proposed hybrid models to combine the advantages of various deep learning techniques. For example, the CNN-long short-term memory (CNN-LSTM) hybrid model integrates CNN for visual feature extraction with LSTM for processing temporal data, improving FER accuracy [24]. Another model, dual-stream CNN [25], processes spatial and temporal information simultaneously through two parallel convolutional streams, making it better at capturing the dynamics of facial expressions. However, despite various approaches, many models still face difficulty recognizing negative expressions. EfficientNet-B3 [26], deep comprehensive multi-patch network [27], spatio-temporal convolutional network [28], and transformer-ResNet [28] have also proven effective in FER [26]-[28]. Adaptive graph convolutional network [29] and temporal FER focus on capturing temporal dynamics [30].

One promising approach to improving FER involves incorporating temporal information in facial images. Recurrent neural networks (RNN), particularly CNN-LSTM, effectively process sequential data and capture the temporal dynamics of facial expressions [31]. By feeding CNN-extracted features as input to LSTM, the model can encode temporal data dynamics while learning visual and temporal patterns. This approach increases classification accuracy compared to FER methods that rely solely on static images. Recent research by Liang *et al.* [31] introduced the deep convolution bidirectional long short-term memory (Bi-LSTM) fusion model, which includes three main components: deep spatial network (DSN) for extracting key features from image locations, deep temporal network (DTN) for monitoring changes over time, and recurrent networks to combine spatial and temporal information for a comprehensive understanding of the situation. This model achieved over 95% accuracy on the extended Cohn-Kanade (CK+) dataset after 10,000 training epochs. However, its reliance on advanced GPUs and large datasets makes it difficult to implement on a larger scale. This study aims to develop a deep residual Bi-LSTM fusion model by incorporating residual blocks from a residual network (ResNet) into the deep convolution Bi-LSTM fusion model to address these issues. This modification aims to reduce the risk of overfitting and improve the model's overall performance, particularly in recognizing both positive and negative expressions more effectively.

2. RELATED WORK

Research on FER has resulted in various models designed to address different challenges and improve accuracy and efficiency. EfficientFace is a model designed with an efficient architecture to overcome limitations in FER and achieve high accuracy [6]. DAN-GAN combines dual attention and generative adversarial networks to enhance image quality and classification of facial expressions [7]. EMO-Net uses deep neural networks to extract important features from facial images and classify them into different expression categories [8]. Transfer leverages the capabilities of transformers in handling long-term dependencies in sequential data, thereby improving FER accuracy [9]. LiteFER is designed for limited-resource devices, enabling fast and accurate FER [10]. However, these models still struggle to detect subtle variations in facial expressions and generalize well across different datasets.

Several other models have been developed to address specific challenges in FER. Occlusion-Aware CNN uses specialized convolutional techniques to handle situations where parts of the face are obscured [11], while Pseudo-3D CNN combines spatial and temporal information to capture the dynamics of facial expressions [12]. ViT applies transformer architecture to divide facial images into small patches and process them in parallel [13]. Emotion transformer and DAN utilize the power of transformers and dynamic attention mechanisms to improve FER accuracy by focusing on key features in the images [14], [15].

Several hybrid models have also been developed to capitalize on various deep learning techniques. HybridNet integrates CNN for visual feature extraction with RNN for temporal data processing, improving both accuracy and efficiency in FER [16]. MobileEmotionNet and TinyFaceNet are designed for mobile devices and IoT with lightweight architectures, allowing fast and accurate recognition even with limited resources [17], [18]. Light-FER focuses on computational efficiency by using compression and optimization techniques in neural networks [19]. Multi-scale feature fusion network integrates information from different image scales to improve FER accuracy [20].

Other effective models for FER include residual masking network, which combines residual and masking techniques to focus on important facial features [21], and Context-Aware FER, which uses contextual information to enhance recognition accuracy [22]. Pyramid attention network leverages pyramid attention mechanisms to capture critical features at various resolution levels [23]. CNN-LSTM hybrid model and dual-stream CNN combine the strengths of CNN and LSTM to process spatial and temporal information simultaneously, making them more effective at capturing the dynamics of facial expressions [24], [25]. Models such as EfficientNet-B3, deep comprehensive multi-patch network, spatio-temporal convolutional network, and transformer-ResNet have also proven effective in FER by combining neural network and transformer techniques to process both visual and temporal data [26]-[28]. Adaptive graph convolutional network and temporal FER focus on capturing temporal dynamics and relationships between features in facial images, thereby improving overall recognition accuracy [29], [30].

3. METHOD

The proposed model, deep residual Bi-LSTM fusion, aims to enhance FER performance by leveraging spatial-temporal information and using residual blocks to prevent overfitting. This method combines the strengths of CNN for feature extraction and Bi-LSTM for capturing temporal dynamics.

3.1. Data preprocessing

This study used three main datasets: real-world affective faces database (RAF-DB), face expression recognition plus (FERPlus), and CK+. The RAF-DB dataset contains approximately 30,000 facial images classified into seven basic emotion categories: surprise, fear, disgust, happiness, sadness, anger, and neutral. Meanwhile, the FerPlus dataset extends the FER2013 dataset with 28,709 training images, 3,589 validation images, and 3,589 test images labeled with eight emotion categories: neutral, happiness, surprise, sadness, anger, disgust, fear, and contempt. The CK+ dataset consists of 593 video sequences from 123 subjects, categorized into seven emotion classes: anger, contempt, disgust, fear, happiness, sadness, and surprise. Images from the CK+ dataset were resized to 48×48 pixels to ensure data consistency, while images from the RAF-DB and FerPlus datasets were resized to 100×100 pixels. All images were normalized to have zero mean and unit variance, which helps to standardize the data and reduce computational complexity. Additionally, various augmentation techniques were applied to increase the diversity of training data and improve the model's generalization ability. These techniques include random rotations within the range of -30 to 30 degrees, horizontal flipping, random cropping, resizing the images back to the appropriate dimensions (48×48 or 100×100 pixels), and random adjustments to brightness, contrast, saturation, and hue. This approach aims to enrich data variety, enabling the model to recognize facial expressions under various conditions more accurately.

3.2. Model architecture

The model architecture can be visualized in Figure 1. The DSN begins with several convolutional layers that extract spatial features from facial images, such as edges, textures, and patterns. Each convolutional layer is followed by batch normalization and rectified linear units (ReLU) activation to introduce non-linearity and stabilize training. Residual blocks learn deeper features and address the vanishing gradient problem.

Each residual block consists of two convolutional layers with 3×3 kernels, followed by batch normalization and ReLU activation. Skip connections are added to allow gradients to flow directly through the network. The features produced by the DSN represent rich spatial information from facial images, which are then used as input to the temporal network. The DTN uses the spatial features extracted by the DSN and accumulates them over time to form a sequence representing the temporal evolution of facial expressions. This sequence of features is fed into Bi-LSTM layers to capture temporal dependencies and motion context. Bi-LSTM processes the sequence in both forward and backward directions, considering past and future information, providing a more comprehensive view of facial expression changes over time. The mathematical equations for LSTM are as (1)-(6):

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (1)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2)$$

$$\tilde{C} = \tan(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (3)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C} \quad (4)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (5)$$

$$h_t = o_t * \tan(C_t) \quad (6)$$

where f_t is the forget gate, i_t is the input gate, \tilde{C} is the candidate cell state, C_t is the cell state, o_t is the output gate, and h_t is the hidden state output. W_f, W_i, W_C, W_o and b_f, b_i, b_C, b_o are the weights and biases that are learned. The outputs from Bi-LSTM in both directions (forward and backward) are combined into a single representation:

$$H_t = [\vec{h}_t; \overleftarrow{h}_t] \quad (7)$$

where \vec{h}_t and \overleftarrow{h}_t are the hidden states from the forward and backward directions, respectively. An attention mechanism is applied to the output of the Bi-LSTM layers to focus on the most relevant temporal features for expression recognition. This attention mechanism computes the attention scores e_t and attention weights α_t as (8) and (9):

$$e_t = v^T \tan(W_h H_t + b_h) \quad (8)$$

$$\alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^T \exp(e_k)} \quad (9)$$

where v and W_h are the learned parameters. The attention output z is:

$$z = \sum_{t=1}^T \alpha_t H_t \quad (10)$$

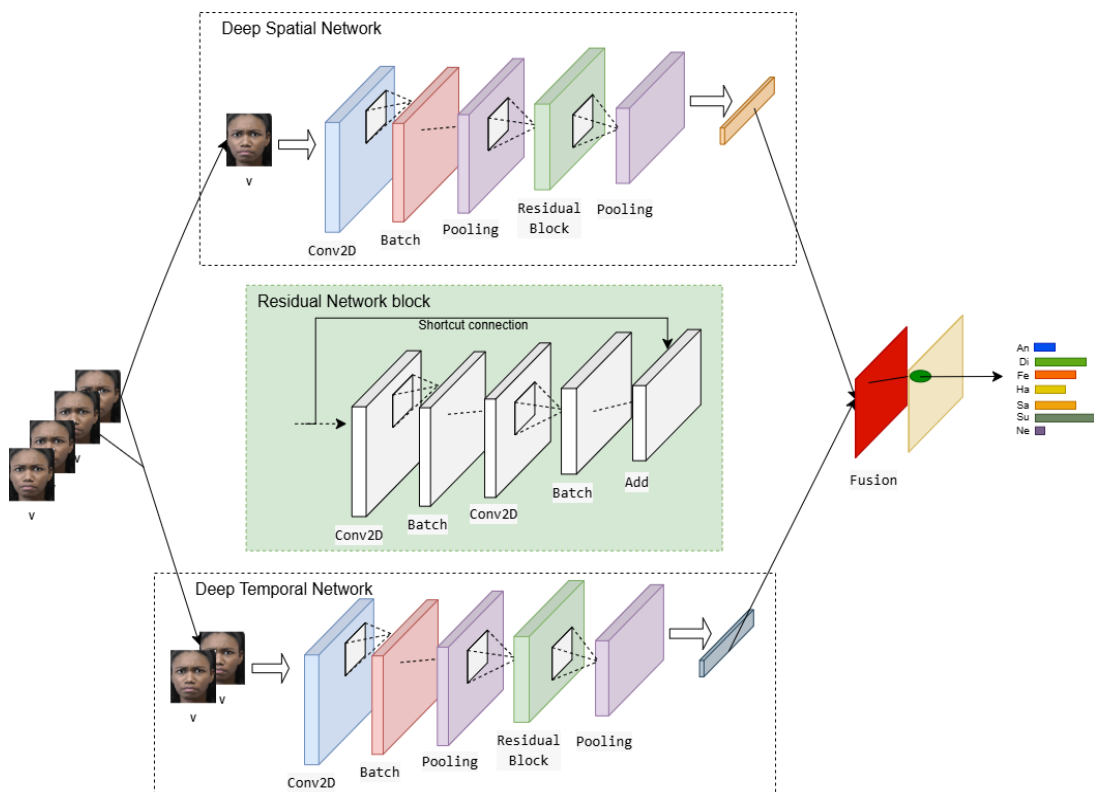


Figure 1. Deep residual Bi-LSTM fusion model

The temporal features from the DTN are then passed through several fully connected layers to map the features to the output space. Dropout layers with a dropout rate of 0.5 are used to prevent overfitting by randomly setting a fraction of the input units to zero during training [31]. A SoftMax activation function is applied in the final layer to produce a probability distribution over the expression classes.

3.3. Training process

The training process involves several steps to ensure the model learns effectively and generalizes well to new data. Categorical cross-entropy loss is used to measure the difference between the predicted and true expression labels, making it suitable for multi-class classification tasks. The Adam optimizer, with an initial learning rate of 0.001, is employed for its adaptive learning rate and robustness across different types of datasets because it is able to adjust learning rates dynamically. This helps in converging quickly and efficiently, making it suitable for handling complex models and diverse datasets [32]. L2 regularization is applied to the weights to prevent overfitting, by adding a penalty to the loss function for large weights. The model is trained using 10-fold cross-validation with a batch size of 4 for 200 epochs [33]. The training process in this study is designed with several key steps to ensure the model learns effectively and generalizes well to new data. Categorical cross-entropy loss is used to measure the difference between the model's predictions and the actual expression labels, making it an ideal choice for multi-class classification tasks. The Adam optimizer is chosen with an initial learning rate of 0.001 due to its advantage in adjusting the learning rate adaptively and its robustness across various types of datasets. Adam's ability to dynamically adjust the learning rate enables the model to converge faster and more efficiently, making it suitable for handling complex models and diverse datasets [32]. To prevent overfitting, L2 regularization is applied to the model weights, adding a penalty to the loss function when the weights become too large. The model is trained using the 10-fold cross-validation technique with a batch size of 4 for 200 epochs [33]. This approach ensures that the model can handle complex data while producing more accurate and reliable results.

3.4. Evaluation

The model's performance is evaluated using several metrics, including accuracy, precision, recall, F1-score, and receiver operating characteristic area under the curve (ROC AUC). Accuracy measures the overall correctness of the model's predictions, while precision represents the proportion of true positive predictions among all positive predictions. Recall indicates the proportion of true positive predictions among all actual positive instances, and the F1-score provides a balanced measure of precision and recall. ROC AUC is used to evaluate the model's ability to distinguish between classes, with higher values indicating better performance. A confusion matrix is generated to visualize the model's performance across different expression classes, showing the number of correct and incorrect predictions for each class. Additionally, the performance of deep residual Bi-LSTM fusion is compared with other state-of-the-art FER models, demonstrating its effectiveness in terms of accuracy, precision, recall, F1-score, and ROC AUC. In summary, the proposed deep residual Bi-LSTM fusion method combines the advantages of CNNs for spatial feature extraction and Bi-LSTMs for capturing temporal dynamics, enhanced with residual blocks to prevent overfitting. This approach aims to achieve high accuracy and robust performance in FER tasks by learning discriminative spatial-temporal information and effectively handling temporal motion context.

The model's performance is evaluated using several key metrics: accuracy, precision, recall, F1-score, and ROC AUC. Accuracy measures the correctness of the model's predictions, while precision indicates the proportion of true positive predictions among all positive predictions. Recall measures how many true positive predictions are made from all actual positive instances, while the F1-score balances precision and recall. ROC AUC assesses the model's ability to distinguish between classes, with higher values indicating better performance. A confusion matrix is generated to visualize the model's performance across different expression classes, showing the number of correct and incorrect predictions for each class.

Additionally, the performance of the deep residual Bi-LSTM fusion model is compared to other state-of-the-art FER models, demonstrating its superiority in terms of accuracy, precision, recall, F1-score, and ROC AUC. Overall, the proposed deep residual Bi-LSTM fusion method combines the strengths of CNNs for spatial feature extraction and Bi-LSTMs for capturing temporal dynamics, further enhanced by residual blocks to prevent overfitting. This approach is designed to achieve high accuracy and robust performance in FER tasks by learning discriminative spatial-temporal information and effectively handling temporal motion context.

4. RESULTS AND DISCUSSION

The training and testing results of the deep residual Bi-LSTM fusion method were evaluated using three main datasets: CK+, RAF-DB, and FerPlus, with evaluation metrics including accuracy, precision, recall, F1-score, and ROC AUC. The CK+ and RAF-DB datasets contain seven emotion classes: Class 0=anger, Class 1=contempt, Class 2=disgust, Class 3=surprise, Class 4=happiness, Class 5=sadness, and Class 6=fear.

Deep residual bidirectional long short-term memory fusion: achieving superior ... (Muhammad Munsarif)

Meanwhile, the FerPlus dataset contains eight emotion classes: Class 0=anger, Class 1=contempt, Class 2=disgust, Class 3=surprise, Class 4=happiness, Class 5=sadness, Class 6=fear, and Class 7=neutral. The distribution of samples for each class across the three datasets varies shown in Table 1. The CK+ dataset has a relatively balanced distribution, with the happiness class having the highest number of samples (69 samples) and the contempt class having the fewest (18 samples). On the other hand, the RAF-DB dataset has a larger overall sample size, with the happiness class having the most samples (7,489 samples) and the fear class the fewest (1,125 samples). The FerPlus dataset shows a more diverse distribution, with the happiness class having the most samples (8,989 samples) and the fear class having the fewest (3,012 samples).

Table 1. Class distribution in datasets

Class	CK+	RAF-DB	FERPlus
0	45	4703	4002
1	18	876	2716
2	59	2820	3515
3	25	3165	3090
4	69	7489	8989
5	28	2716	6077
6	25	1125	3012
7	-	-	6466

Figure 2 shows the loss function graphs for the training and validation processes on three different datasets: CK+, RAF-DB, and FERPlus. Each graph illustrates how the loss values decrease as the number of epochs increases. In Figure 2(a), which represents the CK+ dataset, the initial loss values are quite high but decrease rapidly at the beginning of the training. After around 50 epochs, the loss values stabilize close to zero, indicating that the model consistently reduces prediction errors. Figure 2(b), showing the graph for the RAF-DB dataset, displays a similar pattern. The initial loss values are high but decrease quickly during the first 50 epochs, reaching stability near zero afterward, suggesting that the model also effectively reduces prediction errors on this dataset. Meanwhile, Figure 2(c), depicting the graph for the FERPlus dataset, demonstrates a consistent downward trend. The loss values drop rapidly during the early training phase, then stabilize near zero after approximately 50 epochs. All three graphs show a consistent decline in loss values for training and validation data, indicating that the model learns well and reduces prediction errors over time.

Figure 3 shows the accuracy graphs during training and validation on CK+, RAF-DB, and FERPlus datasets. In Figure 3(a), which represents the CK+ dataset, accuracy is initially low. However, it increases rapidly within the first ten epochs, then stabilizes near the maximum value, indicating that the model quickly achieves high accuracy and maintains it throughout the training process. Figure 3(b), representing the RAF-DB dataset, shows a similar pattern, with accuracy rising sharply within the first ten epochs and stabilizing near the maximum value, suggesting that the model learns quickly and maintains optimal performance. In Figure 3(c), accuracy increases rapidly during the initial training phase for the FERPlus dataset. It stabilizes after about ten epochs, demonstrating the model's ability to achieve and sustain high training and validation data accuracy. All three graphs indicate that the model can quickly improve accuracy during the early training phase and maintain stable, high performance throughout the process, demonstrating effective learning capability and optimal performance across the three datasets.

Figure 4 shows the confusion matrix for classification using the CNN method on three different datasets: CK+, RAF-DB, and FERPlus, illustrating the model's performance in classifying various emotion classes. In Figure 4(a), which represents the CK+ dataset, the model successfully classifies most labels correctly, with Class 4 (happiness) having 185 correct predictions and Class 0 (anger) having 121 correct predictions, indicating good accuracy in recognizing emotions. Figure 4(b), which displays the RAF-DB dataset, shows a similar result with a high number of correct predictions along the diagonal, where Class 5 (happiness) has 3,980 correct predictions, and Class 0 (anger) has 232 correct predictions, demonstrating the model's capability to handle a larger dataset. In Figure 4(c), representing the FERPlus dataset, the model also shows high accuracy, with Class 7 (neutral) having 980 correct predictions and Class 6 (fear) having 967 correct predictions. Overall, the CNN model demonstrates excellent classification performance across the three datasets, with the majority of correct predictions located along the diagonal of the confusion matrix, indicating the model's ability to recognize and classify emotions accurately.

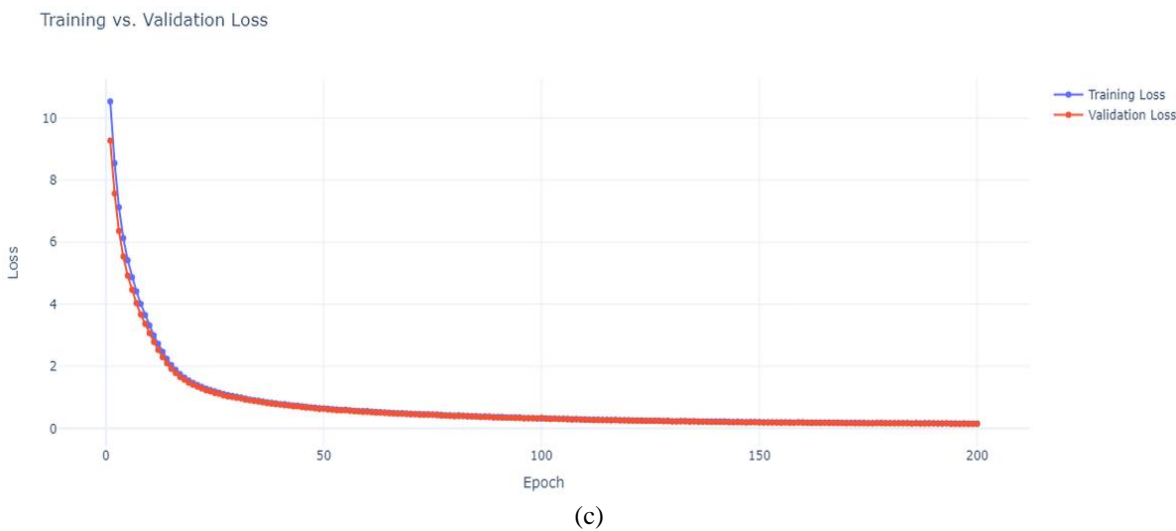
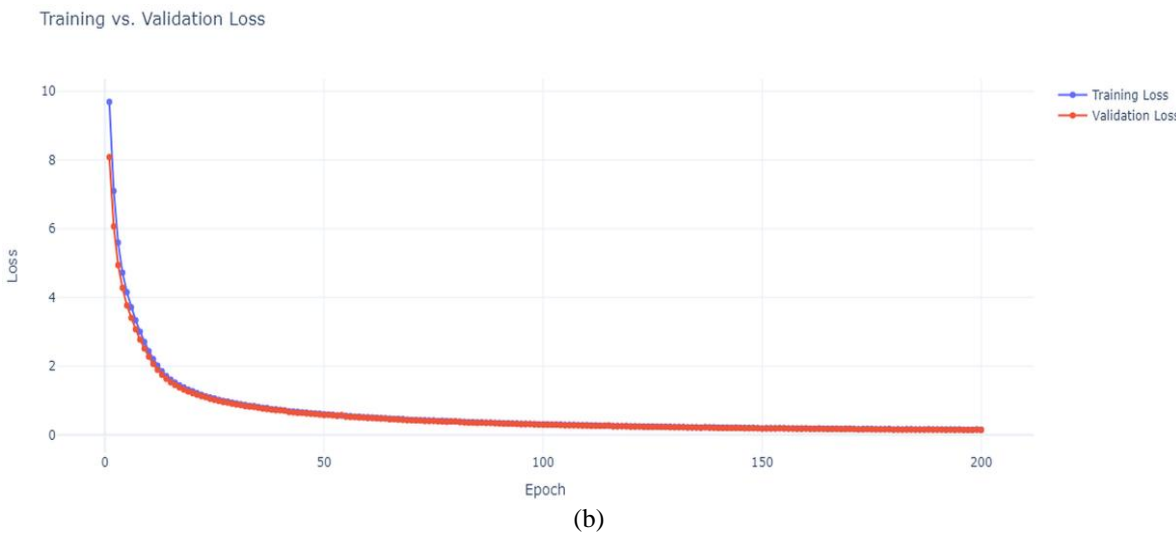
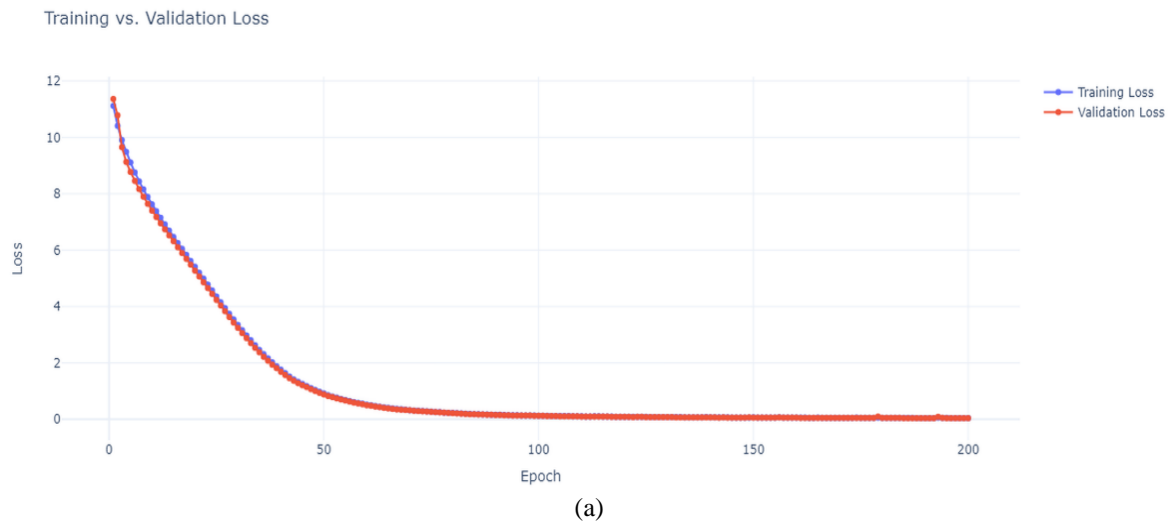


Figure 2. Loss function; (a) Ck+, (b) RAF-DB, and (c) FERPlus

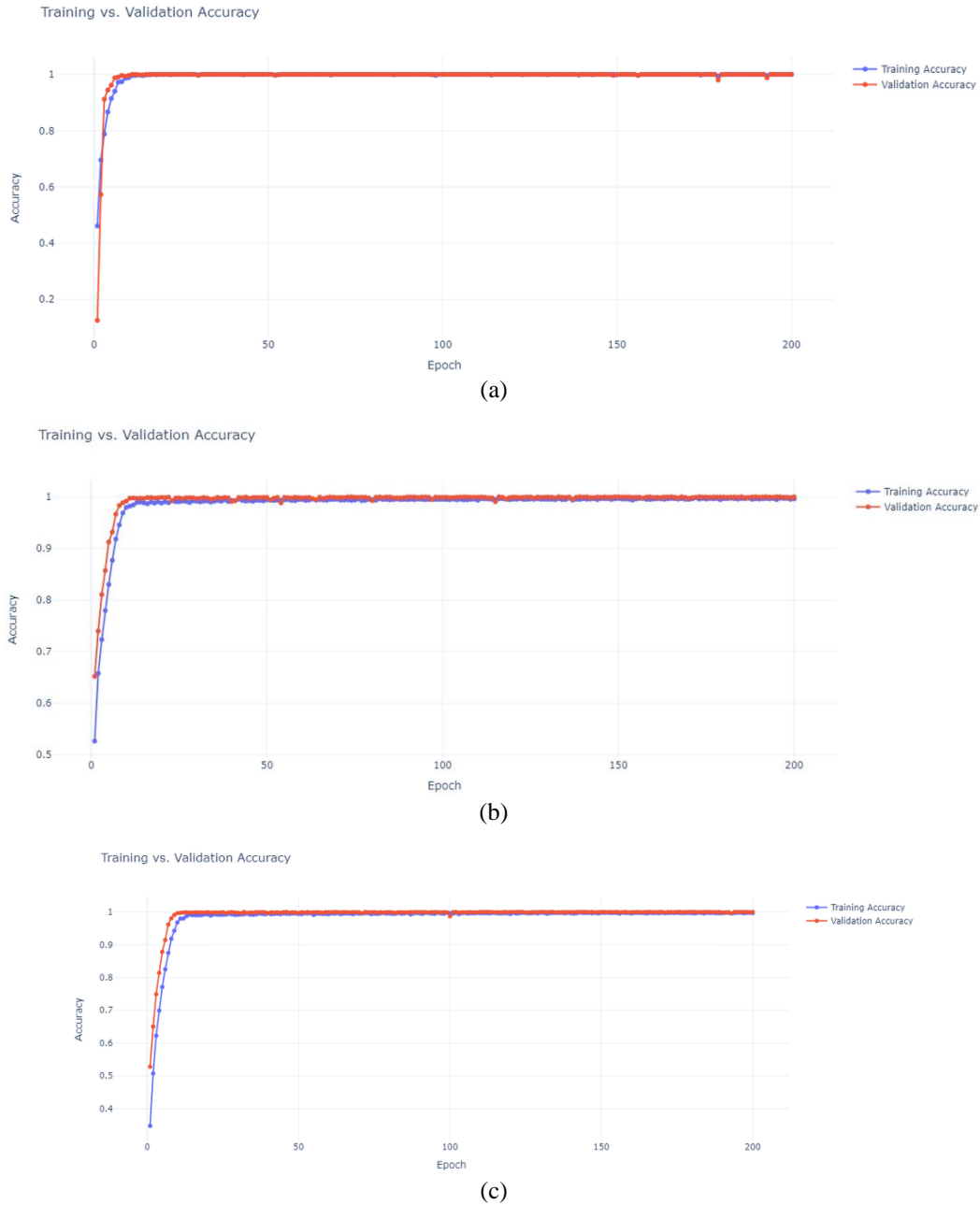
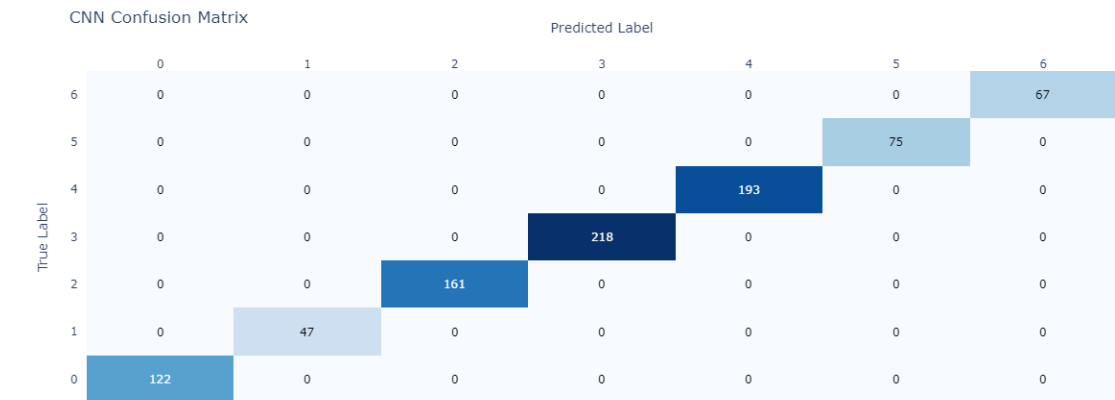
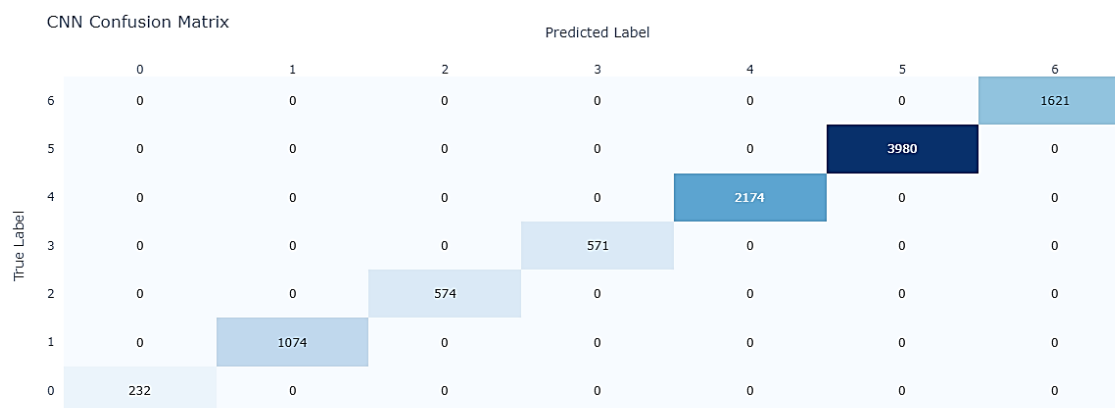


Figure 3. Accuracy function: (a) Ck+, (b) RAF-DB, and (c) FERPlus

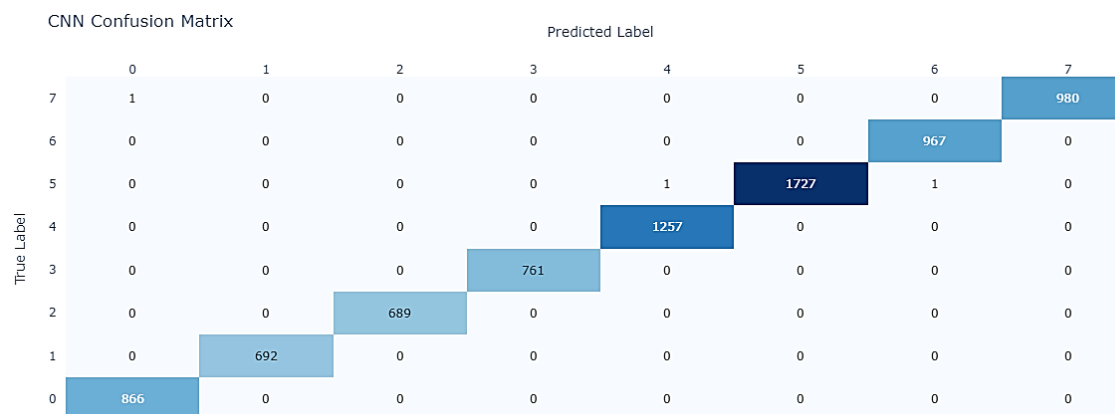
Figure 5 shows the ROC curves for each fold across three different datasets: CK+, RAF-DB, and FERPlus, providing an overview of the model's ability to distinguish between emotion classes based on the true positive rate and false positive rate. In Figure 5(a), which represents the CK+ dataset, the ROC curve demonstrates variations in the model's performance for each emotion class. Some classes, such as Class 1 (contempt) and Class 4 (happiness), have higher AUC values of 0.73 and 0.71, indicating that the model can effectively differentiate between these classes. However, other classes, like Class 0 (anger), show a lower AUC of 0.58, suggesting that the model may struggle to distinguish this emotion from the others. Figure 5(b), which displays the RAF-DB dataset, reveals overall better performance, with nearly all classes having AUC values above 0.6. Class 5 (happiness) achieves the highest AUC of 0.85, demonstrating the model's ability to classify emotions more effectively in this dataset compared to CK+. Meanwhile, Figure 5(c), representing the FERPlus dataset, indicates that the model exhibits excellent performance, with most classes having AUC values above 0.8. Class 1 (contempt) and Class 3 (disgust) achieve the highest AUC values of 0.97 and 0.96, reflecting the model's strong capability in distinguishing these emotions. Overall, Figure 5 illustrates that the model delivers good performance in classifying emotions across all three datasets, with performance variations depending on the complexity and size of the datasets used.



(a)



(b)



(c)

Figure 4. Confusion matrix; (a) Ck+, (b) RAF-DB, and (c) FERPlus

Table 2 presents a comparative analysis of various models and their performance across three datasets: CK+, RAF-DB, and FERPlus. The results indicate that deep residual Bi-LSTM fusion achieved exceptional accuracy, scoring 100% on the CK+ and RAF-DB datasets and 99.96% on FERPlus. In comparison, the highest scores from other models on CK+ include gACNN at 96.40% [11] and paCNN at 97.03% [11], while the SCAN-CNN model reached 97.31% on CK+ and 89.02% on RAF-DB [34], demonstrating competitive performance but not matching the top results of deep residual Bi-LSTM fusion. The IFGAN model achieved 97.52% on CK+ [35], a novel feature decomposition and reconstruction learning (FDRL) achieved 99.54% [36], ViT-SE reported an impressive 99.80% [37], highlighting advancements in model architecture. On the RAF-DB dataset, models such as CIAO and Patt-Lite delivered notable accuracies of 95.05% and 95.55%,

respectively. The TransferFER model performed consistently across both RAF-DB and FERPlus with scores of 90.83% and 90.91% [26]. Overall, the data illustrates that while many models exhibit strong performance, the deep residual Bi-LSTM fusion method is the most effective across all datasets, showcasing its robustness and reliability in FER tasks.

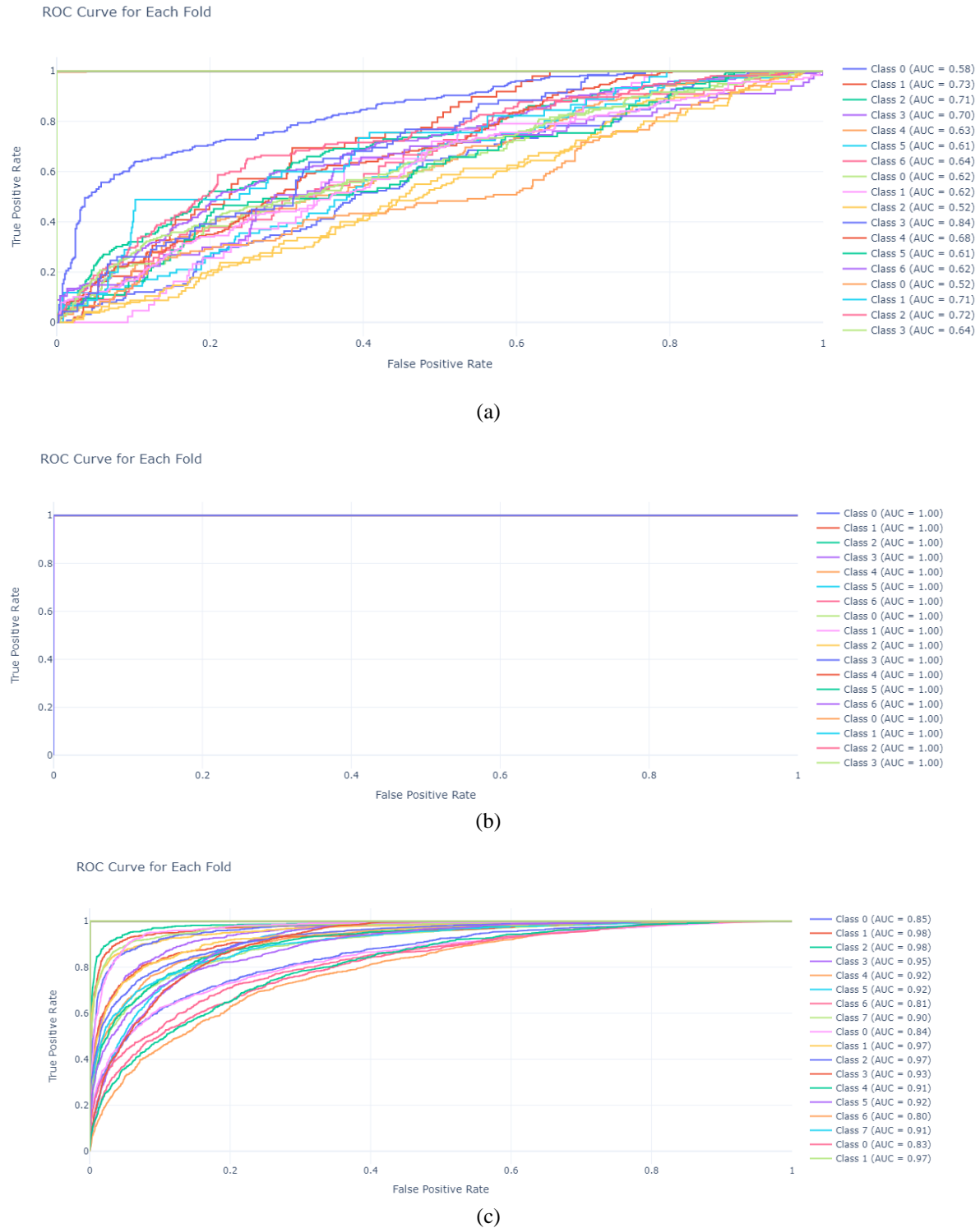


Figure 5. ROC curve; (a) Ck+, (b) RAF-DB, and (c) FERPlus

Table 2. Comparison of accuracy with previous studies based on dataset

Model	Datasets (%)		
	Ck+	RAF-DB	FERPlus
VTFF [9]	-	881.4	88.81
gACNN [11]	96.40	-	-
paCNN [11]	97.03	-	-
TransferFER [26]		90.91	90.83
SCAN-CNN [34]	97.31	89.02	89.42
IFGAN [35]	97.52	-	-
FDRL [36]	99.54	-	-
ViT-SE [37]	99.80	-	-
RAN [38]		86.90	89.16
ARM [39]		90.42	-
Facial chirality [40]		91.20	-
APViT [41]		91.98	90.86
POSTER [42]		92.05	91.62
POSTER++ [43]		92.21	-
CIAO [44]		95.05	95.55
Patt-Lite [45]	100	95.05	95.55
Deep residual BiLSTM fusion	100	100	99.96

5. CONCLUSION

The deep residual Bi-LSTM fusion method has demonstrated superior performance in FER tasks, achieving higher or comparable accuracy to existing methods across the CK+, RAF-DB, and FERPlus datasets. By integrating the strengths of CNN for spatial feature extraction and Bi-LSTM for capturing temporal dynamics, along with the use of residual layers to address the vanishing gradient problem, the model achieved perfect accuracy on CK+ and RAF-DB, and nearly perfect accuracy on FERPlus. These results underscore the model's efficiency and reliability in accurately classifying diverse emotion classes. For future work, we aim to further enhance the model's robustness by incorporating additional real-world datasets, exploring transfer learning techniques, and optimizing the model's architecture to reduce computational complexity without compromising performance. This approach is expected to develop an even more versatile and practical solution for real-time FER applications.

ACKNOWLEDGEMENTS

Author thanks the reviewers for their constructive feedback, colleagues for their insightful discussions, and the research team for their dedication. The author also acknowledges the institutional support and resources provided by Asia e University and Universitas Muhammadiyah Semarang which contributed significantly to the successful completion of this research.

FUNDING INFORMATION

There is no funding for this study.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Muhammad Munsarif	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓
Ku Ruhana Ku-Mahamud		✓		✓		✓				✓		✓		

C : **C**onceptualization

M : **M**ethodology

So : **S**oftware

Va : **V**alidation

Fo : **F**ormal analysis

I : **I**nterpretation

R : **R**esources

D : **D**ata Curation

O : **O**rganizing - **O**rganizing Draft

E : **E**valuation - **E**valuation & **E**valuation

Vi : **V**isualization

Su : **S**upervision

P : **P**roject administration

Fu : **F**unding acquisition

CONFLICT OF INTEREST STATEMENT

There is no conflict of interest.

DATA AVAILABILITY

The datasets used in this study, including FERPlus, RAF-DB, and CK+, are publicly available and can be accessed through their respective sources. The details and access links for each dataset are as follows: i) RAF-DB (<https://shorturl.at/T2yLw>), ii) FERPlus (<https://shorturl.at/E4em0>), and iii) CK+ (<https://shorturl.at/9b1Bn>).




REFERENCES

- [1] Y. Huang, F. Chen, S. Lv, and X. Wang, "Facial expression recognition: A survey," *Symmet. (Basel)*, vol. 11, no. 10, 2019, doi: 10.3390/sym11101189.
- [2] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1113–1133, 2015, doi: 10.1109/TPAMI.2014.2366127.
- [3] M. Sajjad *et al.*, "A comprehensive survey on deep facial expression recognition: challenges, applications, and future guidelines," *Alexandria Eng. J.*, vol. 68, pp. 817–840, 2023, doi: 10.1016/j.aej.2023.01.017.
- [4] N. Sun, Q. Li, R. Huan, J. Liu, and G. Han, "Deep spatial-temporal feature fusion for facial expression recognition in static images," *Pattern Recognit. Lett.*, vol. 119, pp. 49–61, 2019, doi: 10.1016/j.patrec.2017.10.022.
- [5] L. Zhang, B. Verma, D. Tjondronegoro, and V. Chandran, "Facial expression analysis under partial occlusion: A survey," *ACM Comput. Surv.*, vol. 51, no. 2, 2018, doi: 10.1145/3158369.
- [6] G. Wang, J. Li, Z. Wu, J. Xu, J. Shen, and W. Yang, "EfficientFace: an efficient deep network with feature enhancement for accurate face detection," *Multimed. Syst.*, vol. 29, no. 5, pp. 2825–2839, 2023, doi: 10.1007/s00530-023-01134-6.
- [7] M. D. Putro, D. L. Nguyen, and K. H. Jo, "A Dual Attention Module for Real-time Facial Expression Recognition," in *IECON Proc. (Industrial Electronics Conf.)*, 2020, vol. 2020, pp. 411–416, doi: 10.1109/IECON43393.2020.9254805.
- [8] G. Yang, J. S. Y. Ortoneda, and J. Saniie, "Emotion Recognition Using Deep Neural Network with Vectorized Facial Features," in *IEEE International Conf. on Electro Inf. Tech.*, vol. 2018, pp. 318–322, 2018, doi: 10.1109/EIT.2018.8500080.
- [9] F. Ma, B. Sun, and S. Li, "Facial Expression Recognition With Visual Transformers and Attentional Selective Fusion," *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 1236–1248, 2023, doi: 10.1109/TAFFC.2021.3122146.
- [10] E. Bicer and H. Kose, "LITE-FER: A lightweight facial expression recognition framework for children in resource-limited devices," in *18th Int. Conf. on Automatic Face and Gesture Recognition*, 2024, pp. 1–9, doi: 10.1109/FG59268.2024.10581970.
- [11] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion Aware Facial Expression Recognition Using CNN With Attention Mechanism," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2439–2450, 2019, doi: 10.1109/TIP.2018.2886767.
- [12] B. Hasani and M. H. Mahoor, "Facial Expression Recognition Using Enhanced Deep 3D Convolutional Neural Networks," in *IEEE Computer Society Conf. on Comp. Vision and Pattern Recogn. Workshops*, 2017, pp. 2278–2288, doi: 10.1109/CVPRW.2017.282.
- [13] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv*, 2020, doi: 10.48550/arXiv.2010.11929.
- [14] J. Y. Guo *et al.*, "A Transformer based neural network for emotion recognition and visualizations of crucial EEG channels," *Phys. A Stat. Mech. its Appl.*, vol. 603, 2022, doi: 10.1016/j.physa.2022.127700.
- [15] S. M. Saleem, S. R. M. Zeebaree, and M. B. Abdulrazzaq, "Real-life Dynamic Facial Expression Recognition: A Review," in *J. of Physics: Conf. Series*, vol. 1963, no. 1, 2021, doi: 10.1088/1742-6596/1963/1/012010.
- [16] X. Si, D. Huang, Y. Sun, S. Huang, H. Huang, and D. Ming, "Transformer-based ensemble deep learning model for EEG-based emotion recognition," *Brain Sci. Adv.*, vol. 9, no. 3, pp. 210–223, 2023, doi: 10.26599/bsa.2023.9050016.
- [17] N. El Zarif, L. Montazeri, F. Leduc-Primeau, and M. Sawan, "Mobile-Optimized Facial Expression Recognition Techniques," *IEEE Acc.*, vol. 9, pp. 101172–101185, 2021, doi: 10.1109/ACCESS.2021.3095844.
- [18] A. George, C. Ecabert, H. O. Shahreza, K. Kotwal, and S. Marcel, "EdgeFace: Efficient Face Recognition Model for Edge Devices," *IEEE Trans. Biometrics, Behav. Identity Sci.*, vol. 6, no. 2, pp. 158–168, 2024, doi: 10.1109/TBIOM.2024.3352164.
- [19] A. M. Pascual *et al.*, "Light-FER: A Lightweight Facial Emotion Recognition System on Edge Devices," in *Sensors*, vol. 22, no. 23, 2022, doi: 10.3390/s22239524.
- [20] Y. Jiang, S. Xie, X. Xie, Y. Cui, and H. Tang, "Emotion Recognition via Multiscale Feature Fusion Network and Attention Mechanism," *IEEE Sens. J.*, vol. 23, no. 10, pp. 10790–10800, 2023, doi: 10.1109/JSEN.2023.3265688.
- [21] L. Pham, T. H. Vu, and T. A. Tran, "Facial expression recognition using residual masking network," *Proc. - Int. Conf. Pattern Recognit.*, pp. 4513–4519, 2020, doi: 10.1109/ICPR48806.2021.9411919.
- [22] A. Jain, S. Nigam, and R. Singh, "Context-Aware Facial Expression Recognition Using Deep Convolutional Neural Network Architecture," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intell. and Lecture Notes in Bioinform.)*, vol. 14531, pp. 127–139, 2024, doi: 10.1007/978-3-031-53827-8_13.
- [23] Q. Wang, T. Wu, H. Zheng, and G. Guo, "Hierarchical pyramid diverse attention networks for face recognition," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 8323–8332, 2020, doi: 10.1109/CVPR42600.2020.00835.
- [24] Y. Ming, H. Qian, and L. Guangyuan, "CNN-LSTM Facial Expression Recognition Method Fused with Two-Layer Attention Mechanism," *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/7450637.
- [25] H. Q. Khor, J. See, S. T. Liong, R. C. W. Phan, and W. Lin, "Dual-stream Shallow Networks for Facial Micro-expression Recognition," in *Proceedings - Int. Conf. on Image Process., ICIP*, 2019, vol. 2019, pp. 36–40, doi: 10.1109/ICIP.2019.8802965.
- [26] I. N. Alam, I. H. Kartowisastro, and P. Wicaksono, "Transfer Learning Technique with EfficientNet for Facial Expression Recognition System," *Rev. d'Intelligence Artif.*, vol. 36, no. 4, pp. 543–552, 2022, doi: 10.18280/ria.360405.
- [27] A. R. Hazourli, A. Djeghri, H. Salam, and A. Othmani, "Deep Multi-Facial Patches Aggregation Network For Facial Expression Recognition," *arXiv*, 2020, doi: 10.48550/arXiv.2002.09298.
- [28] Z. Yu, G. Liu, Q. Liu, and J. Deng, "Spatio-temporal convolutional features with nested LSTM for facial expression recognition," *Neurocomp.*, vol. 317, pp. 50–57, 2018, doi: 10.1016/j.neucom.2018.07.028.
- [29] T. N. Fatyanosa and M. Aritsugi, "An Automatic Convolutional Neural Network Optimization Using a Diversity-Guided Genetic Algorithm," *IEEE Acces.*, vol. 9, pp. 91410–91426, 2021, doi: 10.1109/ACCESS.2021.3091729.
- [30] P. A. Gavade, V. Bhat, and J. Pujari, "Facial Expression Recognition in Videos by learning Spatio-Temporal Features with Deep




- Neural Networks,” *Proc. IEEE Int. Conf. Image Inf. Process.*, vol. 2021, pp. 359–363, 2021, doi: 10.1109/ICIIP53038.2021.9702545.
- [31] D. Liang, H. Liang, Z. Yu, and Y. Zhang, “Deep convolutional BiLSTM fusion network for facial expression recognition,” *Vis. Comput.*, vol. 36, no. 3, pp. 499–508, 2020, doi: 10.1007/s00371-019-01636-3.
- [32] J. L. Ba and D. P. Kingma, “Adam: A method for stochastic optimization,” *arXiv*, 2015, doi: 10.48550/arXiv.1412.6980.
- [33] R. Manorathna, *K-Fold Cross-Validation Explained In Plain English: For evaluating a model’s performance and hyperparameter tuning*. 2020.
- [34] D. Gera and S. Balasubramanian, “Landmark guidance independent spatio-channel attention and complementary context information based facial expression recognition,” *Pattern Recognit. Lett.*, vol. 145, pp. 58–66, 2021, doi: 10.1016/j.patrec.2021.01.029.
- [35] J. Cai *et al.*, “Identity-Free Facial Expression Recognition Using Conditional Generative Adversarial Network,” *Proc. - Int. Conf. Image Process. ICIP*, vol. 2021, pp. 1344–1348, 2021, doi: 10.1109/ICIP42928.2021.9506593.
- [36] D. Ruan, Y. Yan, S. Lai, Z. Chai, C. Shen, and H. Wang, “Feature Decomposition and Reconstruction Learning for Effective Facial Expression Recognition,” in *Proc. of the IEEE Computer Society Conf. on Comp. Vision and Pattern Recogn.*, 2021, pp. 7656–7665, doi: 10.1109/CVPR46437.2021.00757.
- [37] M. Aouayeb, W. Hamidouche, C. Soladie, K. Kpalma, and R. Seguier, “Learning Vision Transformer with Squeeze and Excitation for Facial Expression Recognition,” *arXiv*, 2021, doi: 10.48550/arXiv.2107.03107.
- [38] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, “Region Attention Networks for Pose and Occlusion Robust Facial Expression Recognition,” *IEEE Trans. Image Process.*, vol. 29, pp. 4057–4069, 2020, doi: 10.1109/TIP.2019.2956143.
- [39] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016, pp. 770–778, 2016, doi: 10.1109/CVPR.2016.90.
- [40] L. Lo, H. Xie, H. H. Shuai, and W. H. Cheng, “Facial Chirality: From Visual Self-Reflection to Robust Facial Feature Learning,” *IEEE Trans. Multimed.*, vol. 24, pp. 4275–4284, 2022, doi: 10.1109/TMM.2022.3197365.
- [41] F. Xue, Q. Wang, Z. Tan, Z. Ma, and G. Guo, “Vision Transformer With Attentive Pooling for Robust Facial Expression Recognition,” *IEEE Trans. Affect. Comput.*, vol. 14, no. 4, pp. 3244–3256, 2023, doi: 10.1109/TAFFC.2022.3226473.
- [42] C. Zheng, M. Mendieta, and C. Chen, “POSTER: A Pyramid Cross-Fusion Transformer Network for Facial Expression Recognition,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops, ICCVW 2023*, 2023, pp. 3138–3147, doi: 10.1109/ICCVW60793.2023.00339.
- [43] J. Mao, R. Xu, X. Yin, Y. Chang, B. Nie, and A. Huang, “POSTER++: A simpler and stronger facial expression recognition network,” *arXiv*, 2023, doi: 10.48550/arXiv.2301.12149.
- [44] P. Barros and A. Sciutti, “CIAO! A Contrastive Adaptation Mechanism for Non-Universal Facial Expression Recognition,” *2022 10th Int. Conf. on Affective Comp. and Intelligent Interac. (ACII)*, Nara, Japan, 2022, pp. 1–8, doi: 10.1109/ACII55700.2022.9953863.
- [45] J. Le Ngwe, K. M. Lim, C. P. Lee, T. S. Ong, and A. Alqahtani, “PAtt-Lite: Lightweight Patch and Attention MobileNet for Challenging Facial Expression Recognition,” *IEEE Access.*, 2024, doi: 10.1109/ACCESS.2024.3407108.

BIOGRAPHIES OF AUTHORS



Muhammad Munsarif    received his Master's degree in Computer Science in 2002 and Ph.D. in 2023 from Dian Nuswantoro University, Semarang, Indonesia. He is now a postgraduate student at the School of Graduate Studies, Asia e University, Selangor, Malaysia. He is also a lecturer in informatics Engineering at Muhammadiyah University, Semarang (UNIMUS). His research interests include computer vision and data science. He can be contacted at email: m.munsarif@unimus.ac.id.



Ku Ruhana Ku-Mahamud    holds a Bachelor's degree in Mathematical Sciences and a Masters degree in Computing. Her Ph.D. degree is in Computer Science. Her research interests include ant colony optimization, pattern classification, and vehicle routing problem. Currently, she is attached to Universiti Utara Malaysia as an Emeritus Professor and Universiti Muhammadiyah Malaysia as a Professor. She can be contacted at email: ruhana@uum.edu.my.